基于学者社交网络的论文与项目关联模型 *

王　柳，汤　庸†，杨佐希，傅城州，毛承洁，毛超丹
(华南师范大学 计算机学院，广州 510631)

摘　要：针对学者社交网络的独特用户，提出一种基于学者社交网络的论文与项目数据的协同关联模型。首先采用两步特征选择方法预处理数据，去除无关和冗余特征，得到影响论文与项目关联的有效特征；然后通过文本向量空间模型 TVSM(text vector space model)计算论文与项目之间的文本相似度，为不同的论文/项目形成推荐集合。通过面向科研人员的社交网络"学者网"数据，实现模型并真实应用于学者网。在线应用情况和用户反馈表明，该模型具有较好的准确性和实用性，可更加充分地挖掘论文与项目之间蕴涵的丰富信息，给用户提供更加高效方便的学术科研管理服务，为分析学术大数据提出了新颖的研究方法。

关键词：社交网络；协同关联模型；特征选择；文本相似度；学者网
中图分类号：TP　　doi: 10.19734/j.issn.1001-3695.2018.10.0820

Association model of paper and project based on scholar social network

Wang Liu, Tang Yong†, Yang Zuoxi, Fu Chengzhou, Mao Chengjie, Mao Chaodan
(*School of Computer Science*, *South China Normal University*, *Guangzhou* 510631, *China*)

Abstract: Considering the unique users of scholars' social networks, this paper proposes a collaborative association model of paper and project data based on scholars' social networks. Firstly, The proposed model uses the two-step feature selection method to preprocess the data, while removing the irrelevant and redundant features. So that the model would obtain the effective features that affect the association between the paper and the project. Then it would adopt text vector space model to calculate the text similarity between the paper and the project. After finishing these, it could form recommendation sets for different papers/projects. Through the social network "SCHOLAT" data for researchers, the model is implemented and applied to SCHOLAT. The online application situation and user feedback show that the model has good accuracy and practicability. Furthermore, it can more fully explore the potential relationship between the paper and the project, provide users with better academic research management services, and propose a novel research method for analyzing the academic big data.

Key words: social network; collaborative association model; feature selection; text similarity; scholat

0　引言

近年来，随着互联网的高速发展，信息数据呈现指数级增长，如何从海量的数据中获取有效的信息数据成为数据挖掘的重要挑战之一[1]。社交网络也是如此，大量的论文、项目等学术成果信息导致学者社交网络出现了信息过载问题[2]，学者用户对于论文与项目的关系挖掘需求也日益激增。其中，最能体现学者科研成果信息的论文和项目，蕴涵了丰富的学者信息资源，这使得它们在学者社交网络中的占据了十分重要的地位。但是目前，用户难以充分的挖掘两者蕴涵的有效信息，关于学者社交网络的论文和项目协同关联模型研究很少。因此，如何针对于学者社交网络的特殊学者用户，为其提供准确、个性化的论文/项目推荐，以便用户能够更好地关联论文与项目，充分挖掘论文与项目之间的信息，成为了一个亟需解决的研究课题。

关联规则挖掘是数据挖掘研究中的一个重要分支,它能帮助用户发现大量数据集中的某种潜在关系[3]。因此，建立论文和项目的协同关联模型对于挖掘其蕴涵的丰富学者信息资源很有必要。在本文中主要研究论文与项目之间的关联模型的构建，方便用户挖掘两者之间的关系。考虑到在论文/项目中存在很多不同特征属性，需要在挖掘信息之前，对数据进行预处理和特征选择。这些特征之中，对于论文与项目关联关系而言，有的是有效特征，有的是无关特征和冗余特征。基于此问题的考虑，在此模型中使用了两步特征选择方法[4]，对论文的特征和项目的特征进行预处理，得到有效的论文特征和项目特征。由于在学者社交网络中，论文与项目的存在大多通过文本格式存储，在借鉴了传统的协同过滤推荐模型的基础上，采用了 TVSM 模型计算论文与项目之间的文本相似度，形成不同论文邻域/项目邻域，由相似度从大到小排序找到目标论文/项目的推荐集合。结合用户的需求输入，形成最终的推荐集合，为用户提供更加准确的、个性化的论文/项目关联选择，最终将协同关联模型可视化展现。通过协同关联模型，可以更加清晰、深入地了解的论文与项目之间的包含信息，发现它们的潜在的关系，并且可从更多不同角

收稿日期：2018-10-30；**修回日期：**2018-11-27　**基金项目：**国家自然科学基金面上项目（61772211）；广东省科技计划项目（2017A040405057，2016B010124008）；广州市产学研协同创新重大专项项目（201704020203）

作者简介：王柳（1995-），女，湖北孝感人，硕士研究生，主要研究方向为社交网络、协同计算；汤庸（1964-），男（通信作者），教授，博导，主要研究方向为社会网络服务、大数据应用、时态数据库等（ytang@m.scnu.edu.cn）；杨佐希（1995-），男，硕士研究生，主要研究方向为社交网络、大数据应用；傅城州（1986-），男，讲师，主要研究方向为社交网络服务、信息安全等；毛承洁（1964-），女，教授级高级工程师，硕导，主要研究方向为电子商务、社交网络；毛超丹（1994-），女，硕士研究生，主要研究方向为社交网络、协同计算.

度进行数据挖掘，为用户提供一种准确、个性化和实用的科研信息管理工具。本文提出的基于学者社交网络的论文与项目数据关联模型真实运用于学者网，通过调查分析用户的反馈情况显示，此模型具有较好的准确性和实用性。

1　相关工作

近年来，基于社交网络的论文推荐研究有很多，黄泳航等人[5]针对学术社交网络特有的社交性，构建了基于社区划分的学术论文推荐模型。在学术社交网络中，通过标签传播划分社区，并依据划分结构在各社区内部的用户之间推荐学术论文。陈国华等人[6]针对于学者网计算机类论文语料库，提出了基于单个词的语义向量计算学术论文文档的语义向量的搜索方案，并真实应用于学者网。汤志康等人[7]提出了一种学术社交平台相似论文推荐算法。该算法首先用 ANSJ 对论文进行分词并统计词条的 TF-IDF，然后通过 Word2Vec 把论文映射到一个高位向量，并使用余弦相似度计算相似度。这些研究只涉及到学术论文的相关搜索推荐，并没有考虑到学者项目也蕴涵着丰富的学者信息，也没有深入研究论文与项目的协同关联。

针对于学者社交网络，在论文/项目中有很多不同的特征属性，在挖掘相关信息之前，需要对数据进行预处理和特征选择。在数据的预处理和特征选择方面，Kira 等人[8]提出了 RELIEF 算法，这是一种经典的基于二分类的特征权重算法。该算法依据各个特征和类别的相关性赋予不同的权重值，权重小于设定阈值的特征会被移除，最后得到各个特征的平均权重。由于 RELIEF 局限于二分类问题，张翔等人[9]通过融合间距最大化和极大熵理论，对于两类数据，多类数据和在线数据，提出了新的 RELIEF 特征加权算法，具有更好的适应性。但是该算法由于赋予了所有和类别相关性高的特征较高的权重值，所以该算法存在不能有效去除冗余特征的局限性。Ding 等人[4]在分析了社交网络中用户特征信息之后，提出了一种融合 RELIEF 算法和 K-means 算法的两步特征选择方法，取得更好的特征集合。实验结果也表明该算法适用于复杂的社交网络数据，具有较好的性能。

为了解决社交网络中存在的信息过载问题，推荐系统应运而生。在推荐算法中，协同过滤算法(collaborative filtering,CF)是个性化推荐系统最为成功的算法之一，该类算法主要通过用户—项目评分矩阵,进行相似度计算，找出目标用户的邻居集合进行推荐。目前主要有基于记忆的协同过滤算法(memory-based CF)[10]和基于模型的协同过滤算法(model-based CF)[11]两种类型。其中，基于记忆的协同过滤算法可以划分为基于用户的协同过滤算法(user-based CF)[12]和基于物品的协同过滤算法(item-based CF)[13]。该类算法往往通过用户—物品评分矩阵，结合相似度算法，计算不同用户/物品的相似度，以此找到目标用户/物品的最相似用户/物品构成最近邻居集合，形成推荐集。在该类协同算法中，相似度计算是关键的步骤之一。作为衡量两个个体之间差异的大小，相似度越高，说明个体间的差异往往较小，依据相似度的推荐的质量也通常越好。在相似度的计算中，度量的方法目前主要有余弦相似性[14]、调整的余弦相似性[15]和 Pearson 相关系数[16]，根据实际的数据情况进行合适的度量方法选择。

综上所述，在本文提出的模型中采用更适合社交网络环境下融合 RELIEF 算法和 K-means 算法的两步特征选择方法进行原始数据的特征选择。在此基础上，考虑到论文和项目的特征属性主要由自然语言文本构成，在借鉴传统协同过滤推荐算法相似度计算和传统的空间向量模型 VSM[17]特征相似度计算的基础上，本文采用了 TVSM 模型计算论文与项目之间的文本相似度，找到目标论文/项目的邻居集合，形成推荐集合并结合用户的个性化需求向用户推荐。

2　模型设计与实现

2.1　特征选择方法

在两步特征选择方法中，RELIEF 算法具有运行效率高，对噪声有容错能力，不受特征交互影响等特点，因此适用于复杂的社交网络数据。特征选取第一步,使用 RELIEF 算法[8]去除特征矩阵中的与项目和论文关联不相关的特征。设有原始的数据集 $D=\{\{x_1,x_2,\cdots,x_m\},\{y_1,y_2,\cdots,y_n\}\}$,原始数据中的两个数据集 $\{x_1,x_2,\cdots,x_m\},\{y_1,y_2,\cdots,y_n\}$ 分别代表论文和项目的特征集合。从训练集 D 中选取一个样本 R ，R 由 p 维向量 $\{x_1,x_2,\cdots,x_p\}$ 组成， p 为特征数， $R(j)$ 为样本 R 的第 j 个特征的值。两个样例 R_1,R_2 关于特征 j 的距离定义如下：

a) 当特征为非数值型变量时

$$diff(R_1(j),R_2(j))=\begin{cases}1, if & R_1(j)=R_2(j)\\ 0, if & R_1(j)\neq R_2(j)\end{cases} \tag{1}$$

b) 当特征为数值型变量时

$$diff(R_1(j),R_2(j))=\left|\frac{R_1(j)-R_2(j)}{\max(j)-\min(j)}\right| \tag{2}$$

其中： $\max(j),\min(j)$ 分别表示特征 j 的最大最小取值。该算法通过找到与 R 同类的最近邻样本 RH 以及与 R 非同类的最近邻样本 RM ，然后依据样本 R 与它的两个最近邻样本在特征 j 上的距离差更新特征 j 的权重，如式（3）所示。

$$W(j)=W(j)-\frac{diff(R_i(j),RH_i(j))}{m}+\frac{diff(R_i(j),RM_i(j))}{m} \tag{3}$$

其中： $W(j)$ 为特征 j 的权重；初始特征权重均为 0； m 为随机抽取样本的次数；i 为抽取的第 i 个样本。通过 m 次的迭代之后，可得到每个特征的平均权重。权重值越大,说明该特征的分类能力更好，越能代表该类别。算法运行结束后，将权重集合 T 按照从大到小排序，依据给定阈值 α 去除权重小于阈值的特征。

通过第一步的 RELIEF 特征选择方法后，过滤了特征选项中的无关特征，但是得到的特征集合仍存在部分的冗余特征。在第二步特征选择中，将通过结合 K-means 聚类算法[18]解决该问题。K-means 是一种基于划分的无监督算法，能够简单、快速地解决聚类问题，对于大数据集的处理，也具有很好的伸缩性和高效性。通过第一步选择得到了过滤了无关特征的特征集 T ，给定需要划分数据集的簇数 k ，首先通过随机选择距离尽可能远的 k 个起始点作为 k 个簇类的质心；然后通过计算数据集中剩余点与各个簇的质心距离远近，将剩余点分配到距离最近的簇类。对于每一个簇类，计算簇类中所有样本的均值作为新的质心，若质心收敛则结束；否则继续迭代计算除去新的质心后剩余点到新质心的距离，以同样的方式选出新的质心，直至最终收敛，或者达到迭代的上限则结束聚类过程，得到最终的聚类结果。同一个簇内的相似度较高，不同簇间的相似度较低。结合第一步中得到的特征权重，删除同簇中权重值较低的冗余特征，得到最终影响项目和论文关联的有效特征。

融合 RELIEF 和 K-means 算法的两步特征选择方法的算法流程如图 1 所示。

融合 RELIEF 和 K-means 的两步特征选择算法的伪代码如下:

两步特征选择算法:

Input: *Training set* $D=\{\{x_1,x_2,\cdots,x_m\},\{y_1,y_2,\cdots,y_n\}\}$;
Sample size m, *Selected Rounds of Sample R*,
Number of features p;
Threshold α; *Feature Weighting Set T*; *Feature Weighting* $W(\ )$;
distance $diff(\ )$;
Number of cluster k; *cluster* $=\{c_1,c_2,\ldots,c_k\}$; *number of* Clustering q
Process:
$T=\varnothing$; $W(\ )=\{0,0,\cdots,0\}$
for $t=1,\ldots,m$:
　select sample R randomly from D
　select the neighbor set H, M of R from the samples of same class and different class respectively
　for $j=1,\ldots,p$:
　　$W(R(j))=W(R(j))-diff(R(j),RH(j))/m+diff(R(j),RM(j))/m$
for $j=1,\ldots,p$:
　if $W(j)\geq\alpha$ then:
　　$T.append(D_j)$
Remove irrelevant features from T and return new features *Fnew*
select k features randomly form *Fnew* as original centroids,
　F is the remaining features that removes the *centroids*
for $t=1,\ldots,q$:
　for f_t in F:
　　for c_k in cluster:
　　　compute the distance between f_t and c_k
　　put the f_t into the closet (the minimum distance) c_k
　update the centroid of every cluster
　if centroids convergence **then**:
　　break;
end
Output *cluster*

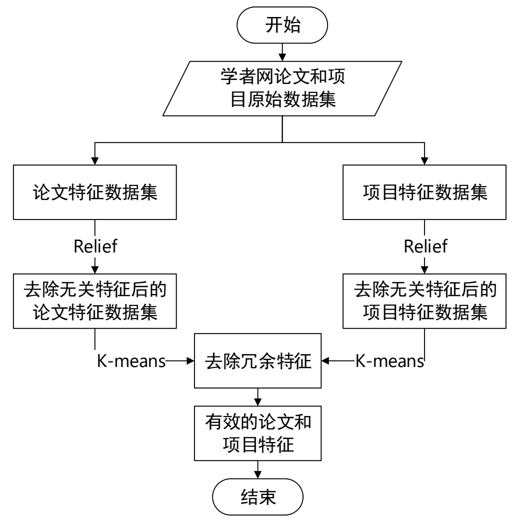


图 1　两步特征选择方法流程

Fig. 1　Two - step feature selection method flow chart

2.2　TVSM 模型

在本文的模型中，为了更好地实现论文和项目的协同关联，需要对目标论文/项目提供更加准确的推荐结果，方便用户更好地从推荐集中选择对应的论文/项目进行关联。去除无关和冗余特征后，考虑到论文信息和项目信息大部分由自然语言组成，本文采用了 TVSM 模型来计算不同文本之间的相似度。首先将论文和项目信息分别进行分词处理，得到论文和项目对应的词组 $D_x=(w_{x1},w_{x2},\cdots,w_{xm}),D_y=(w_{y1},w_{y2},\cdots,w_{ym})$，$m$ 表示词组中的词语总数。然后给每个词组中的词赋予权重值，构成特征向量。其中，权重值通过的 TF-IDF 方法[19]进行计算，计算方法如式（4）所示。

$$TF-IDF(w_i)=tf(w_i)\times\log\frac{n}{df(w_i)} \tag{4}$$

其中：$tf(w_i)$ 表示当前词 w_i 在某个文本中出现的频率；$df(w_i)$ 表示有多少个文本出现了 w_i；n 表示文本总数。由此得到每个文本的单词的 TF-IDF 值，再为每篇文本构建向量模型，利用余弦相似度[14]计算文本之间的相似度,计算方法如式（5）所示。

$$\cos(D_x,D_y)=\frac{\sum_{i=1}^{m}(w_{xi}\times w_{yi})}{\sqrt{\sum_{i=1}^{m}(w_{xi})^2}\times\sqrt{\sum_{i=1}^{m}(w_{yi})^2}} \tag{5}$$

根据相似度的从大到小排序，形成目标论文/项目的最近邻居集合，可以为用户提供选取的推荐集合。

2.3　协同关联模型

论文/项目协同关联模型的框架如图 2 所示。主要组成包括 a）两步特征选择有效的论文和项目特征；b）运用 TVSM 模型计算论文与项目的文本相似度得到相似推荐集；c）结合用户的个性化需求，形成最终目标论文/项目的推荐结果；d）论文与项目关联，系统同步更新。

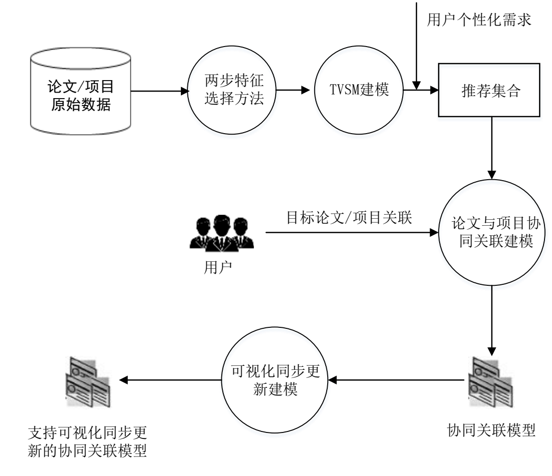


图 2　基于学者社交网络的论文与项目数据协同关联模型

Fig. 2　Association model of paper and project based on scholar social network

本文所提出的协同关联模型，是针对于学者社交网络的特殊学者，为其最具代表信息的学术成果论文和项目提供协同关联。考虑到论文和项目信息本身含有大量的无关和冗余特征，首先通过 RELIEF 算法根据各个特征和类别的相关性赋予特征不同的权重值，移除小于阈值的无关特征；然后通过 K-means 聚类算法依据相似性度量将特征划分为 k 个簇，将簇内权值较低的特征除去，也就是去除冗余特征。通过融合这两种算法的两步特征选择方法进行有效的特征选择，预处理数据。而由于论文和项目的信息基本上是由文本组成，结合文本向量空间模型 TVSM 来计算文本特征之间的相似度。先将论文信息和项目信息分别进行分词处理，通过 TF-IDF 给每个词组中的词赋予权重值，构成特征向量,最后通过利用余弦相似计算文本之间的相似度。通过上述的计算后，按照相似度大小排列，可以得到各个不同的论文/项目的相似推荐集合。为了更好地为用户提供准确、个性化的推荐，在学者网应用实践中，提供了个性化需求输入，以便用户根据自身需求找到更加准确的目标项目/论文。在此基础上，进行论文与项目协同关联模型的建模，用户根据该模型进行论文与项目的协同关联，最终系统将关联结果可视化同步更新。

2.4　协同关联模型应用

本文模型采用学者社交网络—学者网的论文和项目数据进行协同关联模型的构建。学者网(http://www.scholat.com)

是一个面向学者的在线学术信息服务平台，主要提供的功能包括学者在线交流、学术信息管理、文献检索等。以学者网的数据集为例，论文的特征有作者、论文题目、来源、关键字、摘要、类型等十几项之多，项目特征也有 9 项左右，可见，论文和项目各自存在了很多的特征。这些特征之中，对于论文与项目关联关系而言，有的是有效特征，有的是冗余特征。通过两步特征选择选择方法，选择论文数据的“作者”“题目”和“来源”特征和项目数据中的“标题”“参与人员”和“来源编号”特征，并将模型应用与学者网中，在线服务于广大学者用户。模型的图谱如图 3 所示。

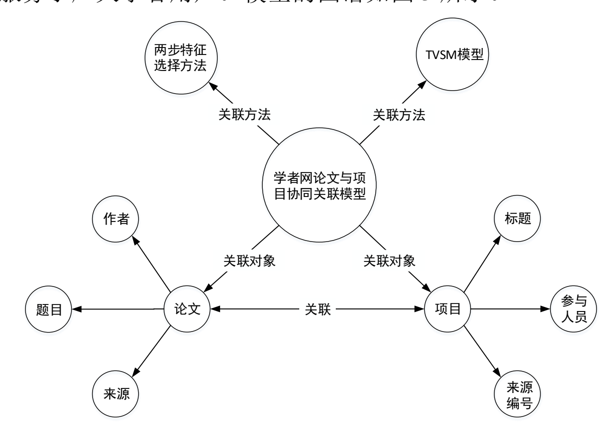


图 3　学者网论文与项目协同关联模型图谱

Fig. 3　Graph of association model of paper and project based on SCHOLAT

在学者网的实际运用中，用户在项目界面的个性化需求的输入框中，根据自己的需求输入具体想要关联的论文相关信息。如图 4 所示，依据用户输入的时态数据管理、时态等具体信息，系统结合用户输入字段与模型所选取的论文的相似推荐集，给出最终的相匹配的推荐列表，推荐结果按时间倒序依次列出。



图 4　结合用户需求与论文相似推荐集的个性化推荐

Fig. 4　Personalized recommendation that combines user requirements and similar recommendation sets of paper

当用户进行论文与项目的协同关联时，通过论文与项目的协同关联模型以及可视化模型进行建模之后，同步更新关联模型的可视化结果，论文界面与项目关联可视化结果如图 5 所示。



图 5　论文界面协同关联项目

Fig. 5　Paper associated with projects

通过上述研究，本文将提出的基于学者社交网络的论文与项目数据协同关联模型最终得以实现并在线应用于学者网，为广大学者用户提供了一种准确、个性化和实用的科研信息管理工具。通过在线问卷调查分别对用户对于推荐结果的满意程度以及用户是否最后选择了关联论文/项目进行了分析，最后的调查结果显示学者网用户对此关联功能是满意的。可以看出，模型的推荐的效果性能较好，能够根据用户的个性化需求推荐出较好的结果，具有较好的实用性。

3　结束语

本文提出了一种基于学者社交网络的论文与项目数据协同关联模型。在该模型中，首先通过融合 RELIEF 和 K-means 算法，对学者网的原始论文和项目数据的无关特征和冗余特征进行过滤；然后采用 TVSM 模型对论文和项目进行文本相似度计算，形成论文/项目的相似推荐集，并结合用户的个性化需求选取更准确的推荐。最终用户从推荐结果选取所需信息进行论文与项目的协同关联，系统同时将论文项目的协同关联同步可视化更新在论文和项目界面。通过学者网站的在线应用情况和用户反馈情况表明，该模型具有较好的准确性和实用性，能更好地帮助学者用户管理科研成果，更深入挖掘论文与项目之间蕴涵地丰富资源与信息。在将来的工作中，可通过这些关联关系进一步研究学者用户间的学术关系，构建用户的知识图谱并进行用户画像，为用户提供更好的服务。

参考文献：

[1] Xu R, Wang S, Zheng X, *et al*. Distributed collaborative filtering with singular ratings for large scale recommendation [J]. Journal of Systems & Software, 2014, 95 (9): 231-241.

[2] Isinkaye F O, Folajimi Y O, Ojokoh B A. Recommendation systems: principles, methods and evaluation [J]. Egyptian Informatics Journal, 2015, 16 (3): 261-273.

[3] 刘军煜，贾修一. 一种利用关联规则挖掘的多标记分类算法 [J]. 软件学报, 2017, 28 (11): 2865-2878. (Liu Junyu, Jia Xiuyi. Multi-label classification algorithm based on association rule mining [J]. Journal of Software, 2017, 28 (11): 2865-2878.)

[4] Ding Rui, Zhu Jia, Tang Yong, *et al*. A novel feature selection strategy for friends recommendation [C]// Proc of IEEE International Conference on Computer Supported Cooperative Work in Design. 2016: 123-128.

[5] 黄泳航，汤庸，李春英，等. 基于社区划分的学术论文推荐模型 [J]. 计算机应用, 2016, 36 (5): 1279-1283. (Huang Yonghang, Tang Yong, Li Chunying, *et al*. Academic paper recommendation model based on community partition [J]. Journal of Computer Applications, 2016, 36 (5): 1279-1283.)

[6] 陈国华，汤庸，许玉赢，等. 基于词向量的学术语义搜索研究 [J]. 华南师范大学学报: 自然科学版, 2016, 48 (3): 53-58. (Chen Guohua, Tang Yong, Xu Yuying, *et al*. Research on academic semantic search

using word vector representations [J]. Journal of South China Normal University: Natural Science, 2016, 48 (3): 53-58.)

[7] 汤志康，李春英，汤庸，等. 学术社交平台论文推荐方法 [J]. 计算机与数字工程, 2017, 45 (2): 221-225. (Tang Zhikang, Li Chunying, Tang Yong, *et al.* Paper recommendation method based on scholar social platform [J]. Computer and Digital Engineering, 2017, 45 (2): 221-225.)

[8] Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm [C]// Proc of the 10th National Conference on Artificial Intelligence.[S.l.]: AAAI Press, 1992: 129-134.

[9] 张翔，邓赵红，王士同，等. 极大熵Relief特征加权 [J]. 计算机研究与发 展, 2011, 48 (6): 1038-1048. (Zhang Xiang, Deng Zhaohong, Wang Shitong, *et al.* Maximum entropy relief feature weighting [J]. Journal of Computer Research and Development, 2011, 48 (6): 1038-1048.)

[10] Bellog, Alejandro N, Castells P, *et al.* Improving memory-based collaborative filtering by neighbour selection based on user preference overlap [C]// Proc of Conference on Open Research Areas in Information Retrieval. 2013: 145-148.

[11] Yin H, Cui B, Li J, *et al.* Challenging the long tail recommendation [J]. Proceedings of the VLDB Endowment, 2012, 5 (9): 896-907.

[12] Bellogin A, Parapar J. Using graph partitioning techniques for neighbour selection in user-based collaborative filtering [C]// Proc of ACM Conference on Recommender Systems. 2012: 213-216.

[13] Pirasteh P, Jung J J, Hwang D. Item-based collaborative filtering with attribute correlation: a case study on movie recommendation [M]// Intelligent Information and Database Systems. [S.l.]: Springer International Publishing, 2014: 245-252.

[14] 董洋溢，李伟华，于会. 基于混合余弦相似度的中文文本层次关系挖 掘 [J]. 计算机应用研究, 2017, 34 (5): 1406-1409. (Dong Yangyi, Li Weihua, Yu Hui. Hierarchical relation mining of Chinese text based on mixed cosine similarity [J]. Journal of Computer Research and Development, 2017, 34 (5): 1406-1409.)

[15] Ma Z, Yang Y, Wang F, *et al.* The SOM based improved K-means clustering collaborative filtering algorithm in TV recommendation system [C]// Proc of the 2nd International Conference on Advanced Cloud and Big Data. [S.l.]: IEEE Computer Society, 2014: 288-295.

[16] Bobadilla J, Ortega F, Hernando A. Recommender systems survey [J]. Knowledge-Based Systems, 2013, 46 (1): 109-132.

[17] Sidorov G, Gelbukh A, Gómezadorno H, *et al.* Soft similarity and soft cosine measure: similarity of features in vector space model [J]. Computación y Sistemas, 2014, 18 (3): 491-504.

[18] Cohen M B, Elder S, Musco C, *et al.* Dimensionality reduction for k-Means clustering and low rank approximation [C]// Proc of the 47th ACM Symposium on Theory of Computing. 2015: 163-172.

[19] 黄承慧，印鉴，侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报, 2011, 34 (5): 856-864. (Huang Chenghui, Yin Jian, Hou Fang. A text similarity measurement combining word semantic information with TF-IDF method [J]. Chinese Journal of Computers, 2011, 34 (5): 856-864.)